UNSUPERVISED CHORD-SEQUENCE GENERATION FROM AN AUDIO EXAMPLE

Katerina Kosta [1,2], **Marco Marchini** [2], **Hendrik Purwins** [2,3]
[1] Centre for Digital Music, Queen Mary, University of London, Mile End Road, London E1 4NS, UK
[2] Music Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain
[3] Neurotechnology Group, Berlin Institute of Technology, 10587 Berlin, Germany
marco.marchini@upf.edu, katkost@gmail.com, hpurwins@gmail.com

ABSTRACT

A system is presented that generates a sound sequence from an original audio chord sequence, having the following characteristics: The generation can be arbitrarily long, preserves certain musical characteristics of the original and has a reasonable degree of interestingness. The procedure comprises the following steps: 1) chord segmentation by onset detection, 2) representation as Constant Q Profiles, 3) multi-level clustering, 4) cluster level selection, 5) metrical analysis, 6) building of a suffix tree, 7) generation heuristics. The system can be seen as a computational model of the cognition of harmony consisting of an unsupervised formation of harmonic categories (via multi-level clustering) and a sequence learning module (via suffix trees) which in turn controls the harmonic categorization in a top-down manner (via a measure of regularity). In the final synthesis, the system recombines the audio material derived from the sample itself and it is able to learn various harmonic styles. The system is applied to various musical styles and is then evaluated subjectively by musicians and non-musicians, showing that it is capable of producing sequences that maintain certain musical characteristics of the original.

1. INTRODUCTION

To what extent can a mathematical structure tell an emotional story? Can a system based on a probabilistic concept serve the purpose of composition? Iannis Xenakis discussed the role of causality in music in his book "Formalized Music, Thought and Mathematics in Composition", where it is mentioned that a fertile transformation based on the emergence of statistical theories in physics played a crucial role in music construction and composition [20].

Statistical musical sequence generation dates back to Mozart's "Musikalisches Würfelspiel" (1787) [8], and more recently to "The Continuator" by F. Pachet [14], D. Conklin's work [3], the "Audio oracle" by S. Dubnov et al. [6] and the "Rhythm Continuator" by M. Marchini and H. Purwins (2010) [13]. The latter system [13] learns the structure of an audio recording of a rhythmical percussion fragment in an unsupervised manner and synthesizes musical variations from it. In the current paper this method is applied to chord sequences. It is related to work such as a harmonisation system described in [1] which, using Hidden Markov Models, it composes new harmonisations learned from a set of Bach chorals.

The results help to understand harmony as an emergent cognitive process and our system can be seen as a music cognition model of harmony. "*Expectation* plays an important role in various aspects of music cognition" [18]. In particular, this holds true for harmony.

2. CHORD GROUPING

Harmony is a unique feature distinguishing Western music from most other predominantly monophonic music traditions. Different theories account for the phenomenon of harmony, mapping chords e.g. to three main harmonic functions, seven scale degrees, or even finer subdivisions of chord groups, such as separating triads from seventh or ninth chords. The aim of this paper is to suggest an unsupervised model that lets such harmonic categories emerge from samples of a particular music style and model their statistical dependencies.

As Piston remarks in [15] (p. 31), "each scale degree has its part in the scheme of tonality, its tonal function". Function theory by Riemann concerns the *meanings* of the chords which progressions link. The term "function" can be used in a stronger sense as well, for specifying a chord progression [10]. A problem arises from the fact that scale degrees cannot be mapped to the tonal functions in a unique way [4] [16] (p. 51-55). In our framework, the function of a chord emerges from its cluster and its statistical dependency on the other chord clusters.

It is considered that the tonic (I), dominant (V) and subdominant (IV) triads constitute the *tonal degrees* since "they are the mainstay of the tonality" and that the last two give an impression of "balanced support of the tonic" [15]. This hierarchy of harmonic stability has been supported by psychological studies as well. One approach involves collecting ratings of how one chord follows from another. As it is mentioned in [11], Krumhansl, Bharucha, and Kessler

used such judgments to perform multidimensional scaling and hierarchical clustering techniques [9]. The psychological distances between chords reflected both key membership and stability within the key; "chords belonging to different keys grouped together with the most stable chords in each key (I, V, and IV) forming an even smaller cluster. Such rating methods also suggest that the harmonic stability of each chord in a pair affects its perceived relationship to the other, and this depends upon the stability of the second chord in particular" [9].

3. METHODOLOGY

The goal of this system is the analysis of a chord sequence given as audio input, with the aim of generating arbitrarily long, musically meaningful and interesting sound sequences maintaining the characteristics of the input sample.

From audio guitar and piano chord sequences, we detect onsets, key and tempo, and group the chords, applying agglomerative clustering. Then, Variable Length Markov Chains (VLMCs) are used as a sequence model. In Figure 1 the general architecture is presented.

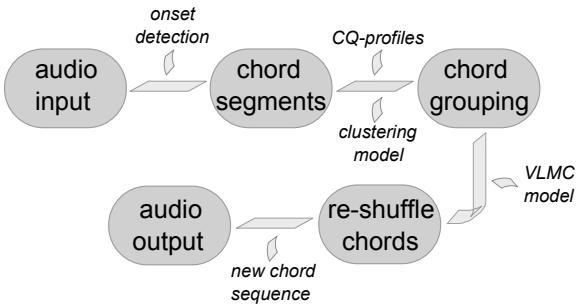

Figure 1. General system architecture.

3.1 Onset Detection

In order to segment the audio into a sequence of chords we employed an onset detection algorithm. Different approaches have been considered since a simplified onset detection method based only on the energy envelope would not be sufficient. After trying a bunch of available algorithms from the literature we found that the *complexdomain* from Aubio [21] was suited for our propose.

A crucial parameter of this algorithm is the *sensitivity* which required an ad hoc tuning. We selected a piano performance of Bach's choral "An Wasserflussen Babylon (Vergl. Nr. 209) in G major - from here on referred as "test -Bach choral" - as a ground truth test set for onset detection. Although with an optimal sensitivity we were still obtaining an incorrect merge of two consecutive segments in the 5.88% of the cases out of a total of 68 segments considered. In Figure 2, the first five segments that were obtained for the test-Bach choral are presented. An example of incorrect merge is shown on the 5th segment, the two consecutive chords of which get still gathered together, as their common notes are still resonating during the passing.

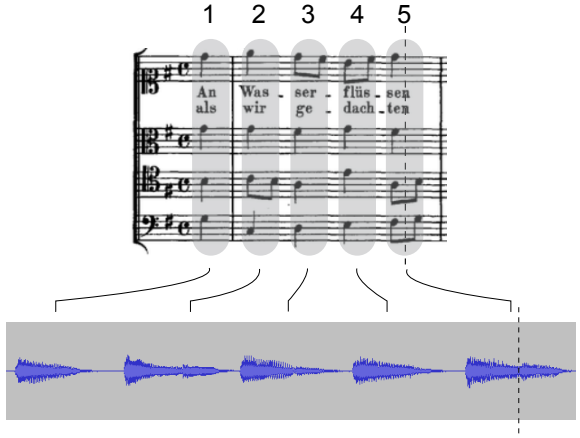

Figure 2. The first 5 segments of the test - Bach choral using Aubio [21] for onset detection. The fifth excerpt should be splitted into two parts -vertical black line- since two different kind of chords are identified and could be used separately.

3.2 Constant Q Profiles and Sound Clustering

From the audio input we extract chroma information based on Constant Q (CQ) profiles, which are 12 - dimensional vectors, each component referring to a pitch class. The idea is that every profile should reflect the tonal hierarchy that is characteristic for its key [2].

The calculation of the CQ profiles is based on the CQ transform; as decribed by Schorkhuber and Klapuri in [19], "it refers to a time-frequency representation where the frequency bins are geometrically spaced and the Q factors which are ratios of the center frequencies to bandwidths, of all bins are equal". This is the main difference between the CQ transform and Fourier transform. In our implementation we have used 36 bins per octave, the square root of a Blackman-Harris window and a hop size equal to 50% of the window size. The CQ profiles are closely related to the probe tone ratings by Krumhansl [17]. Also the system employs a method described by Dixon in [5] for tempo estimation.

In the clustering part, as each event is characterized by a 12-dimensional vector, they can thus be seen as points in a 12-dimensional space in which a metric is induced by the Euclidean distance. The single linkage algorithm has been used to discover event clusters in this space. As defined in [13], this algorithm recursively performs clustering in a bottom-up manner. Points are grouped into clusters. Then clusters are merged with additional points and clusters are merged with clusters into super clusters. The distance between two clusters is defined as the shortest distance between two points, each in a different cluster, yielding a binary tree representation of the point similarities. The leaf nodes correspond to single events. Each node of the tree

occurs at a certain height - level, representing the distance between the two child-nodes (cf. [7] p. 517-557 for details).

Then the *regularity* concept described in [13] is computed for each sequence of each clustering level. Firstly, we compute the histogram of the time differences (CIOIH) between all possible combinations of two onsets. What we obtain is a sort of harmonic series of peaks that are more or less prominent according to the self-similarity of the sequence on different scales. Secondly, we compute the autocorrelation $\text{ac}(t)$ (where t is the time in seconds) of the CIOIH which, in case of a regular sequence, has peaks at multiples of its tempo. Let t_{usp} be the positive time value corresponding to its upper side peak. Given the sequence of m onsets $x = (x_1, \ldots, x_m)$ we define the *regularity* of the sequence of onsets x to be:

$$\text{Regularity}(x) = \frac{\text{ac}(t_{usp})}{\frac{1}{t_{usp}} \int_0^{t_{usp}} \text{ac}(t)dt} \log(m)$$

This regularity is then used to select the most regular level for tempo detection and a small amount of representative levels for the VLMC generation.

In Figure 3, there is a tree representation of the clustering results for the audio test - Bach choral. The system has selected 10 clustering levels, and the cluster hierarchy for the levels 1 - 6 is presented. We have only considered the clusters with more than one element.

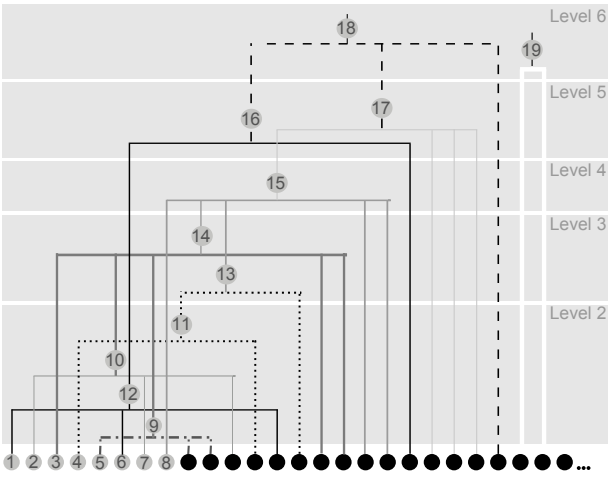

Figure 3. Base line: the clusters generated at Level 1 as circles; the black ones contain one single element.

In Table 1, the clustering results on levels 1 - 4 of the analyzed Bach choral are shown in more detail. It is noticable that we get a *rich* group, containing a large amount of G Major dominant chords.

3.3 Statistical Model for Sequence Generating

Having the segments of the input sound categorized properly, the next step is to re-generate them in a different order than the original one, taking into account that they are not independent and identically distributed, but dependent on the previous segments. For implementing this idea it would be impractical to consider a general dependence of future observations on all previous observations because the complexity of such a model would grow without limit as the number of observations increases. This leads us to consider Markov models in which we assume that future predictions are independent of all but the most recent observations.

A VLMC of order p is a Markov chain of order p, with the additional attractive structure that its memory depends on a variable number of lagged values [12]. This can be evaluated on our system as follows; Let's assume that we have, as an input, two sequences of events - elements of a categorical space having length $\ell = 4$. Be (A,B,C,A) and (B,C,C,D), which are parsed from right to left. As seen in [14], context trees are created where a list of continuations encountered in the corpus are attached to each tree node.

Cluster	# of Elements	Recognition
Level 1:		
cl. 1	3	2 G I, 1 G V
cl. 2	3	1 G I, 1 a V, 1 d IV
cl. 3	2	2 G IV
cl. 4	2	1 G I, 1 G V
cl. 5	10	5 G V, 1 a I, 1 d I, 1 d VI, 1 d V, 1 G I
cl. 6	2	1 G IV, 1 a I
cl. 7	4	1 G II, 1 a V, 1 a I, 1 d V
cl. 8	2	2 G V
Level 2:		
cl. 9	(cl.5)+2	6 G V, 1 a I, 2 d I, 1 d VI, 1 d V, 1 G I
cl. 10	(cl.2+cl.7)+1	1 G I, 2 a V, 1 d IV, 1 G II, 1 a I, 1 d V
cl. 11	(cl.4)+1	2 G I, 1 G V
cl. 12	(cl.1+cl.6)+1	2 G I, 1 G V, 1 G IV, 1 a I
Level 3:		
cl. 13	(cl.11)+1	3 G I, 1 G V
cl. 14	(cl.3+cl.9+cl.10) +2	2 G I, 1 G II, 2 G IV, 6 G V, 2 a I, 2 a V, 2 d I, 1 d IV, 2 d V, 1 d VI
Level 4:		
cl. 15	(cl.8+cl.13+cl.14) + 2	9 G V, 5 G I, 1 G II, 2 G IV, 2 a I, 2 a V, 3 d I, 1 d IV, 2 d V, 2 d VI

Table 1. the clustering results on levels 1 - 4 of the analyzed Bach choral. At the first column, we define each cluster by a number and at the second column we present the number of elements inside that cluster. At the third column we *recognize* these elements and label them based on our score' s harmonic analysis for each one separately (for example: "2 G I" means "2 of the elements are the root of G major" and "5 a V" means "5 of the elements are the dominant of A minor").

The ”continuations” are integer numbers which denote the index of continuation item in the input sequence. In Figure 4, the procedure of the context tree creation based on sequences (A,B,C,A) and (B,C,C,D) is shown, where the index numbers show with which element one can proceed.

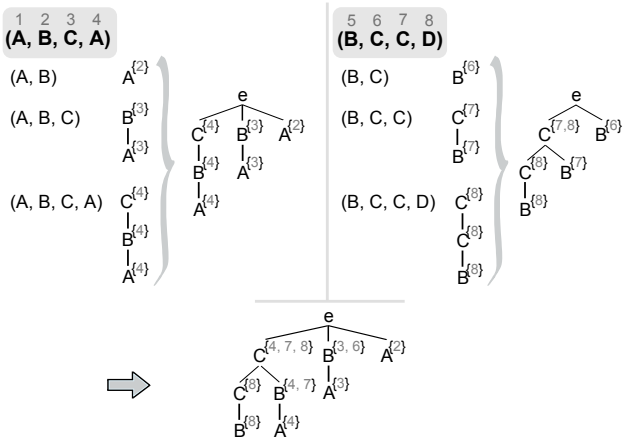

Figure 4. Top left and right: Context trees built from the analysis of the sequences (A B C A) and (B C C D) respectively. Bottom: Merge of the context trees above.

Exploring the final graph in Figure 4, where the trees above are merged, we have all the possible sequence situations, following each path that is created from bottom to up and considering the index number of the first element. For example, if we want to find which is the next element of the sequence (A,B,C), we follow this specific path from the bottom of the tree and then we see the index number of the first element, A, so we take the element with this index number, which is A and the sequence now becomes (A,B,C,A). For ”e” (the empty context) we consider a random selection of any event. Also the length ℓ can be variable.

For the generation we use the suffix trees for all previously selected levels. If we fix a particular level, the continuation indices are drawn according to a posterior probability distribution determined by the longest context found. Depending on the sequence, it could be better to do predictions based either on a coarse or a fine level. In order to increase recombination of blocks and still provide good continuation we employ the heuristics detailed in Section 3.1. in [13] taking into account multiple levels for the prediction.

4. EVALUATION

Five audio inputs have been selected to evaluate the method: a guitar chord sequence based on the song “If I fell in love with you” by the Beatles, a Bach choral played on the piano, part of the “Funeral March” by Chopin, a guitar flamenco excerpt and a piano chord sequence by a non-musician (Examples No.1-5).

The next step was to create generations, using these five different piano and guitar audio inputs followed each one by generations of one minute duration. All the audio examples, some meta data, as well as the generations, and the results of the evaluation are available on the web site [22]. There are two carefully selected generations presented per piece, except for Example No. 5, where there is only one. The following characteristics of the system are assessed: the selected clustering level, the similarity between the input sample and the generation, and how many times an event is followed by another event in the generation that is not the event's successor in the original (i.e. how many “jumps” the generated sound contains).

Since the opinion of a musician rather than an objective measure is a more suitable evaluation measure for the aesthetic value of a generated music sample, a questionnaire for each input and its generations was created and given to five musicians [1] and five non-musicians at ages between 22 and 28. They had to listen to and rate each audio (from 1- “not at all” to 5- “very much”) for their familiarity with the piece and the interestingness of the piece. In addition, the subject had to select the most interesting 10-second parts of it and they had to determine a similarity value comparing two audio examples. Original and the generations were presented without indicating which was which. For Examples 2 and 3 (Bach and Chopin) another question was added, asking to rate how clear the structure of the piece is.

Through the results of this experiment (details in Table 2), we can highlight that only 3% of the responses found the generation example as *not similar* to the original input. Also through the Examples 1, 4 and 5 we notice that 20% of the responses found the generation example more *interesting* than the original and 26% of the responses found the generation example less *interesting*, although the range from the rate of the original one is not big.

In general the cumulative results for the similarity module show small differences between musician's and non-musician's replies. Another measure of comparison between these groups is their response concerning the 10 most interesting seconds; ten groups of overlapping seconds have emerged and seven of these groups were indicated by both musicians and non-musicians.

The comments made by the subjects gave us additional insight into the behaviour of the system. Metrical phase errors have been spotted in the generations of Example No. 4, resulting in rhythmic pattern discontinuities. Some of the musician subjects considered these sections as *“confusing”* and some others as *“intriguing expertise”*. Another important issue is the quality of the generation, in terms of its harmonic structure. A representative comment on Example No.5 is: *“In the second audio (i.e. the Original) I could hear more harmonically false sequences”*.

5. DISCUSSION AND CONCLUSION

The system generates harmonic chord sequences from a given example, combining machine learning and signal processing techniques. As the questionnaire results highlight, the generation is similar to the original sample, maintain-

[1] They are defined as individuals, having at least five years of music theory studies and instrument playing experience.

ing key features of the latter, with a relatively high degree of interestingness.

An important extension of this work would incorporate and learn structural constraints as closing formulae and musical form. Other future work comprises an in-depth comparison of the chord taxonomies generated by the system and taxonomies suggested by various music theorists, e.g. Riemann, Rameau, or the theory of jazz harmony and possibly the experimental verification of such harmonic categories in the brain, e.g. in an EEG experiment.

However, for an automatic music generation system, there remains still a long way to go in order to comply with the idea of music as Jani Christou puts it: "The function of music is to create soul, by creating conditions for myth, the root of all soul".

6. ACKNOWLEDGEMENT

The work of the second author (M. M.) was supported in part by the European project "Social Interaction and Entrainment using Music PeRformance Experimentation" (SIEMPRE, Ref. No. 250026). The work of the third author (H. P.) was supported in part by the German Bundesministerium für Forschung und Technologie (BMBF), Grant No. Fkz $01GQ0850$.

7. REFERENCES

[1] Moray Allan and Christopher K. I. Williams, Harmonising Chorales by Probabilistic Inference, Advances in Neural Information Processing Systems 17, 2005.

[2] J. C. Brown, M. S. Puckette. An efficient algorithm for the calculation of a constant Q transform.,J. Acoust. Soc. Am., 92(5):2698-2701, 1992.

[3] D. Conklin, Music Generation from Statistical Models, Proceedings AISB, p. 30-35, 2003.

[4] C. Dahlhaus, Untersuchungen ber die Entstehung der harmonischen Tonalitt, volume 2 of Saarbrcker Studien zur Musikwissenschaft. Brenreiter-Verlag, Kassel, 1967

[5] S. Dixon, Automatic extraction of tempo and beat from expressive performances, Journal of New Music Research, 30(1):39-58, 2001.

[6] S. Dubnov, G. Assayag, A. Cont, Audio Oracle: A New Algorithm for Fast Learning of Audio Structures, in Proceedings of ICMC, 2007.

[7] R. O. Duda, Peter E. Hartl, D. G. Stork, Pattern Classification (2nd edition), 2001.

[8] K. Jones, Dicing with Mozart: Just a whimsical musicians in the 18th, NewScientist, Physics & Math, 1991.

[9] T. Justus, J. Bharucha, Stevens' Handbook of Experimental Psychology, Volume 1: Sensation and Perception, Third Edition, pp. 453-492, New York: Wiley, 2002.

[10] D. Kopp, On the Function of Function, Music Theory Online, Society for Music Theory, Volume 1, Number 3, May 1995, ISSN: 1067-3040, 1995.

[11] Krumhansl, C.L., Bharucha, J.J., Kessler, E.J. Perceived harmonic structure of chords in three related musical keys. Journal of Experimental Psychology: Human Perception and Performance, vol. 8, pp. 24-36, 1982.

[12] M. Machler, P. Buhlmann, Variable Length Markov Chains: Methodology, Computing and Software, ETH, Research Report No. 104, March 2002.

[13] M. Marchini, H. Purwins, Unsupervised Generation of Percussion Sound Sequences from a Sound Example, MSc thesis, UPF, 2010.

[14] F. Pachet, The continuator: Musical interaction with style, in Proceedings of ICMC (ICMA ed.), pp. 211-218, September 2002.

[15] W. Piston, Harmony, Victor Gollancz Ltd, London, 1959.

[16] H. Purwins, Profiles of Pitch Classes Circularity of Relative Pitch and Key - Experiments, Models, Computational Music Analysis, and Perspectives, Ph.D. Thesis, Berlin Institute of Technology, 2005.

[17] H. Purwins, B. Blankertz, K. Obermayer, A New Method for Tracking Modulations in Tonal Music in Audio Data Format, Proceedings of the IJCNN. vol.6, pp. 270-275, 2000

[18] H. Purwins, M. Grachten, P. Herrera, A. Hazan, R. Marxer, and X. Serra, Computational models of music perception and cognition II: Domain-specic music processing, Physics of Life Reviews, vol. 5, pp. 169-182, 2008.

[19] C. Schorkhuber, A. Klapuri, Constant-Q transform toolbox for music processing, in: 7th Sound and Music Computing Conference, July 2010.

[20] Iannis Xenakis, Formalized Music: Thought and Mathematics in Composition, Bloomington: Indiana University Press, 1971.

[21] http://aubio.org, April 2012.

[22] http://soundcloud.com/chordsequencegenerator , April 2012.

Example 1	Musicians		Non-musicians	
	Familiarity	Interesting	Familiarity	Interesting
Original	2,1,3,1,4	2,2,4 (22-30s),4 (11-16s),3	2,2,3,2,4	3,4(38-42s),4 (22-32s),2,3
Generation 1	2,1,3,1,3	2,2,3,5 (4-12s),2	2,3,2,2,2	4 (1-11s),3,3,2,3
Generation 2	5,1,3,1,2	2,2,3,2,3	3,3,4,2,4	2,5 (48-58s),4 (40-50s),2, 4 (45-55s)
Similarity	Org.-Gen.1	Org.-Gen.2	Org.-Gen.1	Org.-Gen.2
Not similar				
Somewhat similar	++	++	++	+++
Very similar	+++	+++	+++	++

Example 2	Musicians		Non-musicians	
	Familiarity		Familiarity	
Original	4,4,4,5,4		3,3,4,2,5	
	Clearness	Interesting	Clearness	Interesting
Generation 1	4,5,5,3,2	3,5 (30-40s),4 (30-40s),2,1	4,5,4,4,4	4 (1-11s),5 (1-11s), 4 (45-55s),4 (30-40s),3
Generation 2	5,4,3,2,3	1,4 (23-32s),3,3,2	4,4,3,3,3	2,3,4,2,4
Similarity	Org.-Gen.1	Org.-Gen.2	Org.-Gen.1	Org.-Gen.2
Not similar		+	+	
Somewhat similar	+	+	++	+++
Very similar	++++	+++	++	++

Example 3	Musicians		Non-musicians	
	Familiarity		Familiarity	
Original	5,5,4,5,5		4,5,5,5,5	
	Clearness	Interesting	Clearness	Interesting
Generation 1	5,5,5,3,2	5 (0-10s),5 (43-53s),3,3,1	5,5,3,4,3	5 (33-43s),5 (43-48s),3,3 ,5 (30-40s)
Generation 2	5,4,4,2,4	5 (34-44s),4 (43-51s),4,3,2	4,5,3,5,4	5 (17-24s),5 (34-44s),4 (45-52s), 4 (20-30s),4 (40-50s)
Similarity	Org.-Gen.1	Org.-Gen.2	Org.-Gen.1	Org.-Gen.2
Not similar				
Somewhat similar	++	++++		++
Very similar	+++	+	+++++	+++

Example 4	Musicians		Non-musicians	
	Familiarity	Interesting	Familiarity	Interesting
Original	1,2,1,5,2	4 (0-10s),2,4 (34-38s), 4 (28-38s),4 (10-20s)	3,2,3,2,4	4 (1-8s),3,3,1,3
Generation 1	1,2,1,5,2	4 (0-10s),2,3,4 (8-14s), 4 (9-13s)	4,1,4,2,5	3,3,3,1,4 (10-20s)
Generation 2	1,2,1,5,2	1,2,5 (7-15s),3,3	2,1,3,2,5	3,3 (32-42s),4 (45-55s),1,3
Similarity	Org.-Gen.1	Org.-Gen.2	Org.-Gen.1	Org.-Gen.2
Not similar				
Somewhat similar		+++	+	++++
Very similar	+++++	++	++++	+

Example 5	Musicians		Non-musicians	
	Familiarity	Interesting	Familiarity	Interesting
Original	1,2,1,3,4	1,2,2,2,4 (20-30s)	1,1,3,3,3	2,2,2,2,3
Generation	1,2,1,4,4	1,2,3,3,4 (11-16s)	1,1,2,3,2	2,2,3,2,3
Similarity	Org.-Gen.		Org.-Gen.	
Not similar			+	
Somewhat similar	+++++		+++	
Very similar			+	

Table 2. We present the questionnaire responses for Examples 1 - 5; the ratings (from 1 to 5) that both musicians and non musicians have given for each audio thus the rate for similarity comparing specific audio couples are shown. At the *interesting* part, there is a potential mention of the most interesting 10 seconds, in case the response in that section was 4 or 5.